DOCUMENT RESUME

ED 219 424 TM 820 438

AUTHOR Perlman, Carole L.
TITLE Critical Factors in Applying the Value-Added Model: A Monte Carlo Study.
PUB DATE Mar 82
NOTE 39p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Gains; *Age Differences; Early Childhood Education; Evaluation Methods; *Mathematical Models; Measures (Individuals); Performance Factors; *Predictive Validity; Predictor Variables; *Pretests Posttests; Program Evaluation; Reliability; *Statistical Analysis
IDENTIFIERS Monte Carlo Methods; *Value Added Model

ABSTRACT

The value-added model is presented as a method of assessing the impact of treatments in pretest-posttest quasi-experimental designs. The model yields a projection of treatment group performance at posttest time with no special treatment. The difference between the expected result and the actual outcome is "the value added," a measure of treatment effect. No-comparison-group design cases were investigated, and problems and procedures associated with linear and non-linear growth models are discussed. Ages, pretest scores and posttest scores with known means and covariances were generated, and expected growth was calculated from the age and pretest score. The estimated value-added was compared with the actual value-added, and the power of the significance test for the treatment effect was observed. An adequately-specified growth function, sufficient sample size and a reliable, valid test, well-correlated with age and free from ceiling and floor effects, provided an effective model when the measured variable was a function of age or time. Since the model allows for flexibility in the pretest-posttest interval, and does not always require a control group, it was found particularly suitable for early childhood program evaluation. (Author/CM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED219424

CRITICAL FACTORS IN APPLYING THE VALUE-ADDED MODEL
A MONTE CARLO STUDY

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Carole L. Perlman
Chicago Public Schools

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

C. L. Perlman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

INTRODUCTION

The value-added model (Bryk & Weisberg, 1974) provides a method of assessing the impact of treatments in pretest-posttest quasi-experimental designs. The model yields a projection of how the treatment group would have performed at posttest time in the absence of any special treatment; the difference between this expected result and the actual outcome is called the value-added and is a measure of treatment effect. The projection is derived by assuming that in the absence of program effects, test scores are a function of maturational level (and possibly also of background variables such as sex, race, or SES). The regression coefficient of pretest score on age is used as a measure of natural growth rate. Based on that regression and the subject's age at the time of the posttest, an expected posttest score is calculated.

Because of its emphasis on development, the value-added model seems particularly well-suited to early childhood program evaluation, an area in which many other evaluation models have proved difficult to apply. The model allows for a good deal of flexibility in the pretest-posttest interval and has the advantage of not always requiring a control group.

The Model

The following outline of the value-added model is taken from Bryk, Strenio, and Weisberg (1980) and the Huron Institute (1980). Let $Y_i(t)$ be the observed test score for individual i at time t. Then assume:

$$Y_i(t) = G_i(t) + R_i(t) \qquad (1)$$

where $G_i(t)$ represents systematic growth (a function of age or some other time analog), and
$R_i(t)$ is a random noise component whose value is specific to time t.

While $G_i(t)$ can, in principle, be any function of age, it will be assumed to be linear for the sake of simplicity. Thus,

$$G_i = \pi_i\, a_i(t) + \delta_i \qquad (2)$$

TM 820 438

Paper presented at the annual meeting of the American Educational Research Association, New York City, 1982.

where π_i represents the slope (growth rate),

δ_i represents the y-intercept (in effect, time growth started),

t is a constant, and

$a_i(t)$ is the age of subject i at time t.

For the random component, $R_i(t)$, assume that

$E\,[R_i(t)] = 0$,

$\text{var}\,[R_i(t)] = \sigma_R^2$ (equal for all subjects and times),

R_i is independent of t, π_i, δ_i, and

$E\,(R_i R_j) = 0$, $i \neq j$.

At the time of the pretest ($t = t_1$), scores may be represented as

$$Y_i\,(t_1) = \pi_i a_i(t_i) + \delta_i + R_i(t_1) \qquad (3)$$

In the absence of a treatment effect, at the posttest ($t = t_2$), we have

$$\begin{aligned} Y_i\,(t_2) &= \pi_i a_i(t_2) + \delta_i + R_i(t_2) \\ &= G_i(t_1) + \pi_i \Delta_i + R_i(t_2) \end{aligned} \qquad (4)$$

where $\Delta_i = a_i(t_2) - a_i(t_1)$, the time span between pre- and posttest for individual i, and

$\pi_i \Delta_i$ represents expected growth between pre- and posttest due solely to natural maturation.

In the presence of a treatment, the posttest may be represented as

$$Y_i(t_2) = G_i(t_1) + \pi_i \Delta_i + v_i + R_i(t_2) \qquad (5)$$

where v_i, the value-added by the treatment, is the amount an individual's growth is changed by the treatment between t_1 and t_2. The value-added is assumed to be a random variable with

$E(v_i) = \mu_v$, and

$\text{var}(v_i) = \sigma_v^2$.

Another way of stating the model is

intervention effect = observed growth - expected growth in the absence of an intervention.

The observed mean growth is $\bar{Y}_2 - \bar{Y}_1$, the difference between the pretest and posttest scores. The expected mean growth in the absence of an intervention is

$$E(Y_2 - Y_1 | v = 0) = E(\Sigma \pi_i \delta_i)/n \qquad (6)$$
$$= \mu_\pi [\Sigma(\Delta_i/n)] = \mu_\pi \bar{\Delta}$$

If est. (μ_π) is unbiased, an unbiased estimate of μ_v is

$$\text{est. } (\mu_v) = v = \bar{Y}_2 - \bar{Y}_1 - \mu_\pi \bar{\Delta}. \qquad (7)$$

An unbiased estimate of μ may be obtained by regressing pretest on age:

$$\text{est. } (\mu_\pi) = \frac{\Sigma (a_{i1} - \bar{a})(Y_{i1} - \bar{Y}_1)}{\Sigma(a_{i1} - \bar{a})^2} \qquad (8)$$

Ordinary least squares (OLS) can be used to estimate μ_π if it is assumed that age is independent of π and δ. For a fixed Δ, the variance of v is:

$$\text{var}(v) = (\sigma_v^2 + 2\sigma_R^2)/n + \Delta^2\{\text{var}[\text{est.}(\mu_\pi)] - \sigma_\pi^2/n\} \qquad (9)$$

where $\text{var}[\text{est.}(\mu_\pi)] =$

$$\frac{\Sigma (a_{i1} - \bar{a}_1)(a_{i1}^2\sigma_\pi^2 + \sigma_\delta^2 a_{i1}\sigma_{\pi\delta} + \sigma_R^2)}{[\Sigma(a_{i1} - \bar{a}_i)^2]^2} \qquad (10)$$

Since the variance components in Equations 9 and 10 are usually not known, Bryk, Strenio, and Weisberg propose the use of the jackknife technique (outlined in Mosteller & Tukey, 1977) to provide both a test statistic for v and a standard error for use in forming confidence intervals. Jackknifed estimates of the value-added are calculated as follows:

Let est. $[\mu_{\pi(all)}]$ be the OLS coefficient computed in Equation 8 from the whole data set and let v_{*i} be the coefficient computed if the i^{th} observation is removed from the data set. Then for each individual, a pseudo-value, v_{*i}, is calculated as

$$v_{*i} = Y_i(t_2) - Y_i(t_1) - \text{est.}(\mu_{\pi *i})\Delta \qquad (11)$$

where

$$\text{est.}(\mu_{\pi *i}) = n\ \text{est.}[\mu_{\pi(all)}] - (n-1)\ \text{est.}[\mu_{\pi(i)}] \qquad (12)$$

The v_{*i}, which are in effect jackknifed regression coefficients (see Equation 8), are then treated as if they were data points. Their mean, $\bar{v}_*$, is an unbiased estimate of μ_v, and their standard error can be used to calculate a t-statistic with n - 1 degrees of freedom (Miller, 1974). Bryk and Weisberg (1976) suggest using the ratio of actual gain to expected gain as an intuitively appealing measure of program effect. This ratio, however,

requires a measurement scale with a zero point and equal intervals, and would not generally be invariant under transformation.

Bryk, Strenio, and Weisberg (1980) suggest that weighted least squares might yield more precise estimates of the value-added than ordinary least squares in cases where there are unequal error variances across different levels of age, e.g., when ceiling or floor effects are present. A basic assumption of ordinary least squares is that error variances are equal for all values of the independent variable; if this assumption is violated, the resulting regression coefficient estimates will still be unbiased, but they will no longer have the smallest variance of any unbiased linear estimate (Draper & Smith, 1981; Rao & Miller, 1971). After appropriate transformations are performed, least squares solutions can be obtained and the resulting parameter estimates will have smaller variance than OLS estimators.

Other Growth Models

The value-added model as described in the preceding section uses a simple linear growth curve. Strenio (1977) points out that even if the actual long-term growth curve is nonlinear, a linear approximation is likely to be adequate over a relatively short time interval. Bryk and Weisberg (1976) nevertheless indicate that the growth model can assume any desired degree of complexity and suggest that biometric growth models might be employed as part of the value-added strategy.

Bryk, Strenio, and Weisberg (1980) suggest that the assumption of linear growth seems inappropriate in cases where the test exhibits floor or ceiling effects. For example, in the situation where students score near the test ceiling on the pretest, the amount of gain from pretest to posttest might be assumed to be roughly proportional to the initial distance from the ceiling score. Using data from a day care evaluation study, the authors examined several classes of functions whose first derivative (growth rate) is related to distance from the asymptote (ceiling). Bryk, Strenio, and Weisberg ultimately adopted a negative exponential model for systematic growth:

$$G(t) = K\,(1 - \exp\,[\pi a(t) - \delta]) \qquad (13)$$

where

a(t) is the child's age and
K is "some value slightly beyond the celing [sic] value of the test" (p. 27).

In order to simplify the analysis, the experimental data can be transformed in such a way that the individual growth curves become approximately linear. For example, from Equation 14,

$$-\ln [1 - G(t)/K] = \pi a(t) + \delta \quad (14)$$

and

$$G^*(t) = -\ln [1 - G(t)/K] \quad (15)$$

can be defined as a linear individual growth model in the transformed metric. The authors suggest applying the transformation in Equation 15 to the pre- and posttest scores and then proceeding with the value-added analysis as usual.

Threats to the Validity of the Value-Added Model

Some possible difficulties with the value-added model are discussed by the developers of the model and are summarized by Haney (1980). First, strong assumptions are made concerning the nature of children's growth, both in regard to the shape of the growth curve and the types of inferences which may be drawn.

Next, there is the implicit assumption that there does exist

> a natural growth metric corresponding to a developmental dimension whose absolute level for each subject is increasing over time. Many psychometric measures that are actually used are not of this type. The standardization that is typically performed in IQ test scores--forcing the mean and variance to be constant over time--is a classic illustration (Bryk & Weisberg, 1977, p. 959).

It must further be assumed that one does in fact know the shape of the growth curve (linear, exponential, etc.). An incorrectly specified growth model could result in serious under- or overestimates of the value-added. Attempts should be made to correctly specify the age-pretest relationship.

Another possible difficulty is that the value-added model uses cross-sectional data to draw longitudinal inferences (in particular, expected growth in the absence of treatment effects). This requires that one accept what has been called the stable-universe assumption (e.g., Kodlin & Thompson, 1958), namely that the growth of an individual child is independent of the cohort group to which he belongs. For

example, it is assumed that children born in September, 1979, will grow at the same rate as children born six months later. Consequently, the value-added model assumes that there are no historical trends (e.g., the introduction of Sesame Street) that would cause children born at different times to have different growth patterns. Given fairly homogeneous groups, this assumption might be reasonable, unless selection effects are also coming into play, e.g., if the youngest children selected for the program are particularly bright and if the oldest ones are somewhat slower than their peers. In mathematical terms, this would violate OLS assumption of the independence of π and δ, growth rate and starting point. Such selection effects might be minimized by dropping the oldest and youngest children from the analysis.

Another potential problem with the value-added model is assumption that the relationship between age and test score will be the same at pretest and posttest times. This assumption must be made if one is to extrapolate from the pretest data to obtain posttest no-treatment expectations. Strenio (1977) suggests that it is best to keep the extrapolation range small relative to the range of ages in the pretest sample, so that the value-added estimates are as accurate as possible.

Since the value-added model without a comparison group bears some resemblance to Cook and Campbell's (1979) One-Group Pretest-Posttest Design (one of the three weakest quasi-experimental designs), it is useful to compare the internal validity threats to which each of the models is susceptible. Although all five of the validity threats listed below are potential problems with the One-Group Pretest-Posttest Design, the value-added model may serve to overcome some of them:

1. History: Since something other than the treatment could result in a value-added greater than zero, history is a potential internal validity threat to the value-added model.
2. Statistical Regression: Since members of the treatment group are not selected on the basis of pretest score, and since pretest score is not directly used to obtain no-treatment expectations, statistical regression does not appear to be a problem.
3. Maturation: Maturation would not be threat to the value-added model unless the growth curve were incorrectly specified.
4. Testing: When the same test is used more than once, the effects of repeated testing represent a potential validity threat.

5. Instrumentation: This change in the measuring instrument between the pretest and posttest would be likely to cause difficulties only if the test had floor or ceiling effects that were not taken into account when the growth function was specified.

In summary, the value-added model requires some strong assumptions, but it does overcome some of the internal validity threats to which the One-Group Pretest-Posttest Design is subject.

Applications of the Value-Added Model

Although the true experiment is the preferred design in terms of the validity of inferences which can be drawn, it is often not feasible to use that design in the social sciences (Cook & Campbell, 1979). Despite the limitations discussed in the preceding section, Haney (1980) suggests that the value-added model may be appropriate for evaluating the effects of early childhood Title I programs on the development of skills or attributes that show a natural development with age. The major advantage of the model is that it does not necessarily require a control group, although control group data could be used to rule out validity threats by checking the degree to which the model's assumptions were met. A second advantage is that the value-added model provides for great flexibility in the pretest-posttest time interval. It should be noted that use of the value-added model is not strictly limited to preschool program evaluation; it could also be applied in other areas where there is a fairly definite starting point and a pattern of growth over time, e.g., language acquisition.

Problems to Be Investigated

Although the properties of the three approved models for evaluating second- through twelfth-grade Title I programs have been extensively investigated, relatively little work has been done on the value-added model; The purpose of this study is to further investigate the properties and the robustness of the value-added model, particularly in the case of the no-comparison-group design. Several questions are explored. The first three involve the assumption of linear growth:

1. What is the effect of sample size on the significance test for v, the value-added (treatment effect)? While the standard error of

v (calculated using the jackknife) becomes small as the sample size increases, the model might in practice be applied with n as small as 20-30. Thus, it would be desirable to know how the power of the significance test for v is affected when the sample size is varied.

2. How strong must the relationship between test score and age (or test score with age and background variables taken together) be in order for the value-added model to be successfully applied? As the correlation between age and test score decreases, the variance of the estimates of v should increase, so that when H_0: v = 0 is tested, the power is decreased, i.e., the estimate of v becomes less efficient.

3. What happens when the internal test reliability is taken into account? This is related to question 2, since unreliability would contribute to lower correlations between age and test score. This would be a particular problem with tests designed for very young children, since those tests often have relatively low reliability (Haney & Gelberg, 1979).

The other two questions deal with nonlinear growth models:

4. What happens if the underlying growth model is assumed to be linear when it is actually nonlinear? In particular, how are estimates of v affected when the growth function is negative exponential or allometric in form?

5. What is the effect of using weighted least squares to estimate growth rate when the actual growth model is nonlinear? Bryk, Strenio, and Weisberg (1980) suggest the possibility of using generalized least squares rather than ordinary least squares. In cases where there are floor or ceiling effects, error variances may differ considerably for different ages and weighted least squares could provide more efficient estimates of the value-added than one could obtain using OLS.

METHOD

The problems listed in the preceding section were investigated with Monte Carlo simulation, using FORTRAN programs and subroutines from the International Mathematical and Statistical Library (IMSL, 1980).

In the current study, ages, pretest scores and posttest scores with known means and covariances were randomly generated and expected growth was calculated from the age and pretest score. The estimated value-added was compared with the actual value-added and the power of the significance test for v was observed. and investigating the power of that test.

Notation, Definitions, and Assumptions

Let Y_1 = pretest score. Y_1 is distributed normally with mean zero and variance one.

Let X = age at pretest. X is distributed normally with mean zero and variance one.

Let τ = the true test score and ε = the error component of the test score. τ and ε are assumed to be independent. According to classical test theory assumptions:

total test score variance
= true score variance + error variance, so that

$$\sigma_y^2 = \sigma_\tau^2 + \sigma_\varepsilon^2 = 1 \text{ since Y has unit variance.}$$

Let α = test reliability, the ratio of true score variance to total test variance:

$$\alpha = \sigma_\tau^2/\sigma_y^2 = \sigma_\tau^2$$

since Y has unit variance. Let Δ, the interval between the pre- and posttest, = 2.

Linear Growth Models

The design is completely crossed with the correlation between age and true score, the test reliability, the sample size, and the value-added each taking on several different values (see below). There are 100 replications per cell. The following parameter values were chosen because they represent a range of values that might actually be found in practice.

Let $r_{X\tau}$, the correlation between age and true score assume the values .3, .7, .9.

Let α, the test reliability, take on the values .3, .5, .7, .9, 1.0.

Let n, the sample size, take on the values 30, 120.

Let v, the actual value-added, be normally distributed with variance 1.0 and mean μ_v = .1, .25, .5.

The following is an outline of the procedures used in the part of the study dealing with linear growth:

1. Age-true score pairs (X,τ) were generated with age normally distributed with mean zero and variance one, and τ normally distributed with mean zero and variance equal to α, the test reliability. The IMSL subroutine GGNSM was used to generate bivariate standard normal variables X and Y with covariance $(X,Y) = r_{X\tau}$, where $r_{X\tau}$ is the correlation between age and true score. True scores τ were obtained by applying the transformation $\tau = \sqrt{\alpha}\ Y$.
2. Pretest and posttest error terms ε with mean zero and variance $1 - \alpha$ were obtained by applying the transformation $\varepsilon = \sqrt{1 - \alpha}\ \varepsilon'$ to the standard normal variables generated using the IMSL subroutine GGNNL.
3. Pretest scores Y_1 were calculated by adding true score and error terms: $Y_1 = \tau + \varepsilon$, where τ is $(0,\alpha)$ and ε is $N(0, 1 - \alpha)$. Since τ and ε are independent, Y_1 is $N(0,1)$.
4. Using GGNML, additional standard normal variables were generated. These were transformed to obtain the individual treatment effects (values-added), v_i, by adding μ_v. The values-added were distributed normally with mean μ_v and variance one.
5. Posttest scores were calculated such that $Y_2 = \tau + \pi\Delta + v + \varepsilon_2$, where the growth rate, π, is the ordinary least squares regression coefficient of pretest score on age and ε_2 is normally distributed with mean zero and variance $(1 - \alpha)$.
6. Removing the i^{th} observation from the data set, pseudo-values $\mu_{\pi*i}$ and v_{*i} were calculated as described in the introduction. Treating the v_{*i} as if they were data points, the following were obtained: $\bar{v}_*$, an unbiased estimate of μ_v; $s_v^2 = \Sigma\ (v_{*i} - \bar{v}_*)^2\ /\ (n - 1)$; and standard error of v_*, $s_{\bar{v}} = s_v\ /\ \sqrt{n}$.
7. The hypothesis H_0: $v = 0$ was tested at $\alpha = .01$, $.05$ and $.10$ using the statistic $\underline{t} = \bar{v}_*/s_{\bar{v}}$. Critical values for t were obtained using the IMSL subroutine MDSTI.
8. Bias and mean squared error (MSE) were calculated, where bias $= \bar{v}_* - \bar{v}_{actual}$ and MSE $= s_{\bar{v}}^2 + (\text{bias})^2$

Data generation, the analyses described above, and calculation of various other statistics used to measure the adequacy of the simulation, were performed using FORTRAN programs. The programs were run with the FORTRAN H Extended compiler on an IBM 370 computer; One hundred replications were generated for each cell in the design. Double precision arithmetic was used whenever possible in order to maximize accuracy of the results. Each program produced 50 replications of all cells and was run twice in succession. The seeds that were used for the random number generators, GGNSM and GGNML, in the initial run were those that were used in the IMSL documentation. At the end of each run, the current

values of the seeds were printed and then inserted as starting values for the seeds in the next run.

Nonlinear Growth Models

A completely crossed design was used with sample size assuming two values, 30 and 120. The size of the mean actual value-added took on one of three levels, .1σ, .25σ, or 0.5σ, where σ is the standard deviation of posttest scores, 16.0931). The standard deviation of the value-added assumed two values (either equal to, or half of, the posttest standard deviation before a treatment effect was added on). There were 100 replications in each of the 12 cells.

9. Normally distributed ages were generated with mean zero and variance one.

10. Pretest scores, f(age) + ε were calculated, where f is a nonlinear function of age and ε is a normally distributed random error term. Nonlinear growth functions were obtained by fitting curves to the sum of subscale scores of the Chicago EARLY Assessment which was administered to 5,955 children between ages 36 and 58 months as part of the 1980-1981 ESEA Title I program evaluation. The test has a maximum possible score of 107, which is derived by adding individual subtest scores. Two of these curves are linearizations of allometric and negative exponential functions.

$$-\ln (1 - Y/107) = -.06957952 \text{ (age)} - 1.626376 \quad \text{(negative exponential) and}$$

$$\ln (Y) = .9700575 [\ln (\text{age})] + .6355037 \quad \text{(allometric).}$$

The actual functions used in the simulation were the nonlinear forms of the above equations:

$$Y = 107 [1 - \exp(-.06957952 \text{ age} + 1.626376)] + \varepsilon$$

and

$$Y = \exp (.9700575 \text{ age} + .6355037) + \varepsilon.$$

Age was transformed to age in months with a mean of 48 and a standard deviation of 5.5. The normally distributed error term ε was added after exponentiation rather than before, so that the simulated data would conform as nearly as possible to the OLS assumption of normally distributed error terms.

11. Posttest scores were calculated as in the previous section, except that expected growth equals f(age at posttest) - f(age at pretest). The pretest-posttest interval, Δ, is eight months. Because of the constraint that the test score cannot exceed 107 (the maximum

possible score or ceiling), it was sometimes necessary to reduce the posttest score and value-added so that the posttest score would not be greater than 107.

12. The estimated growth rate (μ_π) was calculated using both ordinary least squares and weighted least squares (using age as a weight).
13. Using both ordinary and weighted least squares estimates of the growth rate, the analysis proceeded as outlined in steps 6-8 above.

RESULTS

Linear Growth Model

An analysis of variance performed on mean squared error (MSE) showed significant main effects for sample size, age-true score correlation, and test reliability (see Tables 1 and 2). In general, MSE decreased (i.e., accuracy of estimation increased) as sample size, age-true score correlation, and test reliability increased. For example, when sample size increases from 30 to 120, mean MSE is reduced from 0.19 to 0.04. Increasing test reliability from 0.3 to 1.0 causes the mean MSE to drop by one-half from 0.16 to 0.08. Similarly, when the age-true-score correlation increases from 0.3 to 0.9, mean MSE is reduced from 0.14 to 0.09. The size of the actual value-added had no effect on the accuracy with which v was estimated (see Table 2); the mean MSE was 0.12 for all values of mean actual value-added. Since MSE is a function of the standard error of est.(μ_v), it is hardly surprising that increasing the sample size results in a smaller MSE. Increasing test reliability and age-true score correlation serve to reduce the "error" component of the variance of est.(μ_v). Thus, the precision with which the value-added is measured becomes greater when either test reliability or age-true score correlation is increased.

The ANOVA also yielded several significant interaction effects: sample size by age-true score correlation; sample size by reliability; reliability by age-true score correlation; and sample size by age-true score correlation by reliability. Cell means illustrating the two-way interactions are plotted in Figure 1. The significant interaction of sample size and age-true score correlation is reflected by the fact that although MSE decreases as the correlation increases, the decline in MSE is much more pronounced when N = 30 than when N = 120 (from .24 to .15 from N = 30 and from .05 to .03 for N = 120). Similarly, as test reliability increases, MSE decreases much more rapidly at the smaller sample size than at the larger one. The significant interaction between test reliability and age-true score correlation is explained by the fact that

Table 1

Analysis of Variance of Mean Squared Error
(Linear Growth Function)

Source of Variation	SS	DF	MS	F	Significance
Sample size (SAMSIZ)	47.960	1	47.960	>9999.999	.001
Age-true score correlation (COR)	3.438	2	1.719	416.433	.001
Test reliability (REL)	9.630	4	2.407	583.198	.001
Size of actual value-added (ACTUALV)	0.002	2	0.001	0.213	NS
SAMSIZ by COR	1.348	2	0.674	163.277	.001
SAMSIZ by REL	3.634	4	0.909	220.088	.001
SAMSIZ by ACTUALV	0.006	2	0.003	0.719	NS
COR by REL	0.589	8	0.074	17.836	.001
COR by ACTUALV	0.019	4	0.005	1.162	NS
REL by ACTUALV	0.060	8	0.008	1.823	NS
SAMSIZ by COR by REL	0.326	8	0.041	9.862	.001
SAMSIZ by COR by ACTUALV	0.019	4	0.005	1.127	NS
SAMSIZ by REL by ACTUALV	0.059	16	0.007	1.789	NS
COR by REL by ACTUALV	0.037	16	0.002	0.561	NS
SAMSIZ by COR by REL by ACTUALV	0.038	16	0.002	0.575	NS
Error	37.781	8910	0.004	- - -	- - -

Table 2

Treatment Group Means

(Linear Growth Model)

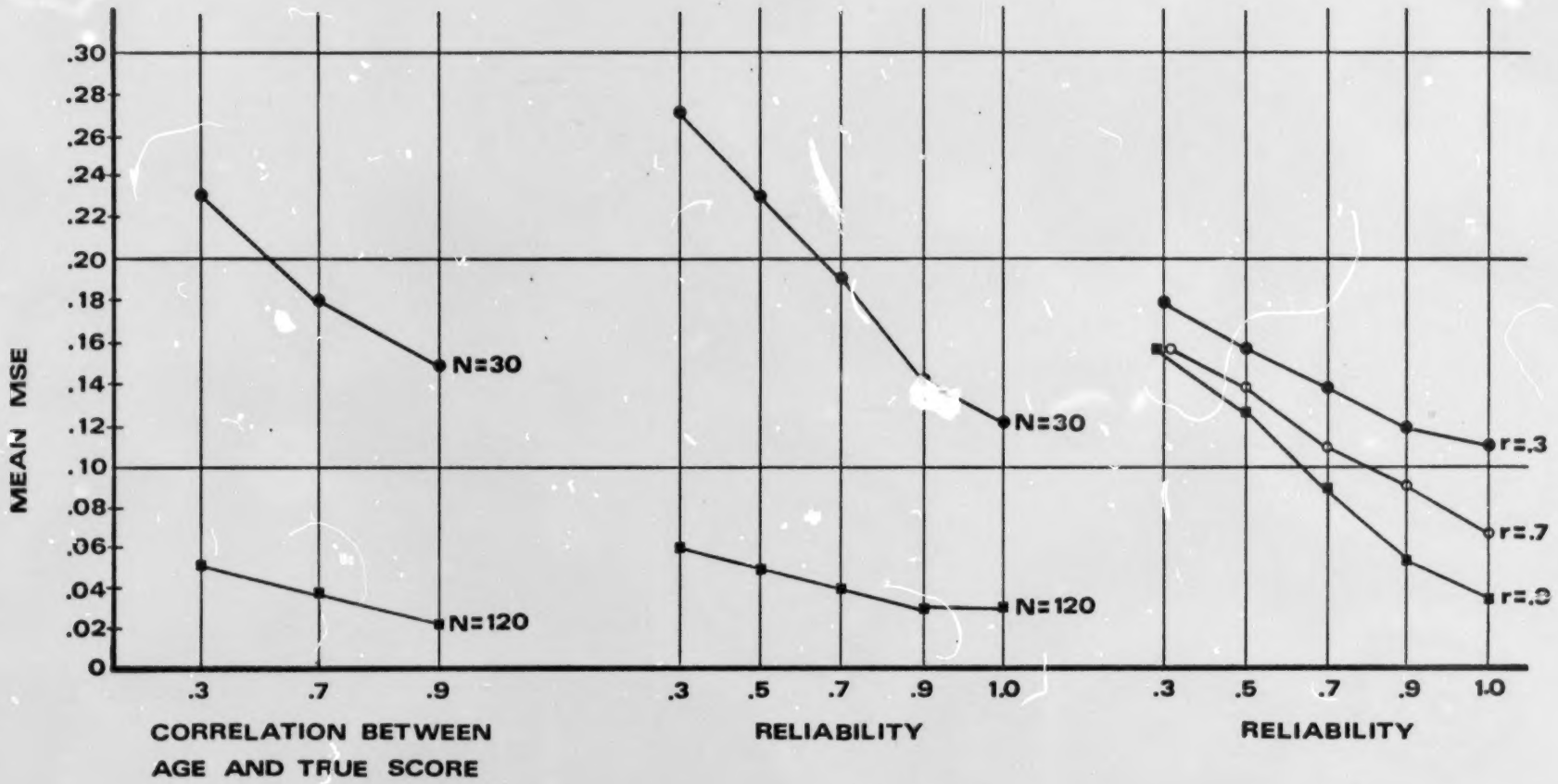
Total Population Mean	0.12 (N=9000)	
Sample Size:	30 0.19 (N=4500)	120 0.04 (N=4500)

Test Reliability:	.3 0.16 (N=1800)	.5 0.14 (N=1800)	.7 0.11 (N=1800)	.9 0.09 (N=1800)	1.0 0.08 (N=1800)

Age-True Score Correlation	.3 0.14 (N=3000)	.7 0.11 (N=3000)	.9 0.09 (N=3000)

Mean Actual Value-Added	.10 0.12 (N=3000)	.25 0.12 (N=3000)	.50 0.12 (N=3000)

Significant Two-Way Interactions from MSE ANOVA



SAMPLE SIZE BY CORRELATION | **SAMPLE SIZE BY RELIABILITY** | **RELIABILITY BY CORREATION**

– Figure 1 –

although MSE decreases as test reliability increases, the amount of that decline becomes greater as the age-true score correlation increases; as test reliability increases, mean MSE drops from .18 to .11 when r = .3, compared to a drop from 0.16 to 0.04 when r = .9.

The previously described phenomena concerning the relationship between MSE, sample size, test reliability, and age are also reflected in the significant three-way interaction among sample size, test reliability and age-true score correlation (see Figure 2). As test reliability increases, MSE declines at a more rapid rate at the higher levels of age-true score correlation and at the lower value of sample size. In general, it appears that variations in test reliability and age-true score correlation have a much greater impact on MSE when the sample size is 30 than when it is 120.

Related to accuracy of estimation of v is the power of the significance test for est.(μ_v). Table 3 gives marginal values of power; and Tables 4 and 5 give the observed power of a one-tailed significance test for est.(μ_v) at the .10, .05, and .01 levels of significance. Graphs are also presented in Figure 3 for the case in which test reliability = 1. Like MSE, power is related to sample size, age-true score correlation, and test reliability; power tends to increase when any of those factors increase. In addition, power depends on the size of the actual value-added, since it is generally easier to detect a large treatment effect than a small one.

On the whole, the power of the significance test for est.(μ_v) is fairly low, particularly when the sample size is small. When α = .05 and N = 30, power was generally less than .10; it exceeded .25 in only two of the 45 cells (mean actual value added = 0.5σ, age-true score correlation = .9, and test reliability = 0.9 or 1.0). In those instances, the power was .33 and .46. At N = 30, variations in test reliability, age-true score correlation, and significance level make relatively little difference in power, especially for detecting a treatment effect of 0.1σ or 0.25σ. Even for the largest treatment effect, .50σ, power exceeds .50 only at α = .10 and when the age-true score correlation is at least .9 and the test reliability is also at least .9.

The situation is somewhat better for N = 120, at least when the treatment effect is relatively large and the test reliability and age-true score correlation are high. However, power is still extremely low for detecting small treatment effects. When the value-added is .50σ, power is relatively good (i.e., above .80) if α = .10; for α = .05, power is fairly low (generally under .25) unless the test reliability is

Table 3

Marginal Values of the Power of Significance Test for v at $\alpha = .01, .05, .10$

(Linear Growth Model)

	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
Total Population	.10	.20	.28
Sample Size			
N = 30	.02	.07	.13
N = 120	.18	.33	.42
Test Reliability			
.3	.05	.15	.23
.5	.08	.17	.24
.7	.08	.18	.26
.9	.13	.24	.31
1.0	.15	.26	.34
Age-True Score Correlation			
.3	.05	.15	.22
.7	.10	.19	.27
.9	.15	.26	.33
Mean Actual Value-Added			
.10	.00	.01	.03
.25	.02	.10	.19
.50	.27	.49	.60

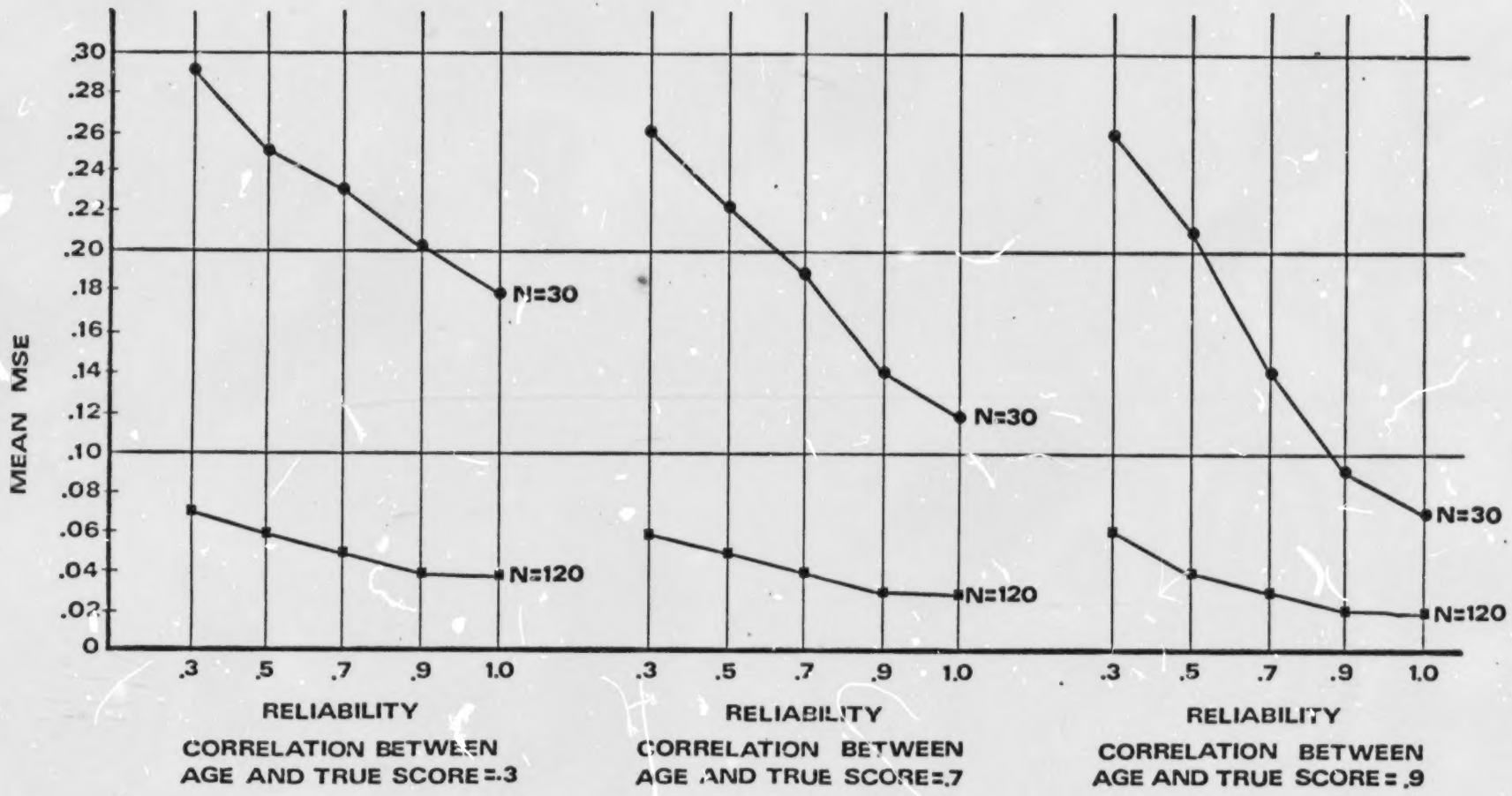
Table 4
Power of Test for v = 0 for N = 30
(Linear Growth Model)

Mean Actual Value-Added	Reliability	Correlation	α = .10	α = .05	α = .01
.10		.3	.01	.01	.00
	.3	.7	.02	.00	.00
		.9	.00	.00	.00
		.3	.00	.00	.00
	.5	.7	.00	.00	.00
		.9	.02	.01	.00
		.3	.00	.00	.00
	.7	.7	.00	.00	.00
		.9	.04	.02	.00
		.3	.01	.00	.00
	.9	.7	.02	.00	.00
		.9	.07	.02	.01
		.3	.00	.00	.00
	1.0	.7	.00	.00	.00
		.9	.06	.01	.01
.25		.3	.08	.01	.00
	.3	.7	.07	.04	.00
		.9	.05	.03	.00
		.3	.06	.03	.00
	.5	.7	.12	.07	.00
		.9	.07	.05	.01
		.3	.02	.01	.00
	.7	.7	.08	.04	.00
		.9	.09	.01	.01
		.3	.02	.00	.00
	.9	.7	.07	.03	.02
		.9	.09	.05	.00
		.3	.03	.02	.00
	1.0	.7	.06	.04	.00
		.9	.16	.03	.02
.50		.3	.22	.10	.00
	.3	.7	.21	.07	.03
		.9	.25	.13	.03
		.3	.20	.12	.01
	.5	.7	.24	.10	.00
		.9	.26	.17	.06
		.3	.18	.07	.01
	.7	.7	.18	.12	.03
		.9	.35	.24	.04
		.3	.20	.11	.02
	.9	.7	.34	.25	.05
		.9	.52	.33	.13
		.3	.20	.05	.01
	1.0	.7	.40	.22	.03
		.9	.61	.46	.19

Table 5
Power of Test for $\bar{v} = 0$ for N =120
(Linear Growth Model)

Mean Actual Value-Added	Reliability	Correlation	$\alpha = .10$	$\alpha = .05$	$\alpha = .01$
.10		.3	.04	.02	.00
	.3	.7	.03	.00	.00
		.9	.04	.01	.00
		.3	.02	.01	.00
	.5	.7	.04	.02	.00
		.9	.08	.03	.00
		.3	.01	.00	.00
	.7	.7	.04	.00	.00
		.9	.06	.01	.00
		.3	.00	.00	.00
	.9	.7	.06	.02	.00
		.9	.12	.06	.00
		.3	.03	.00	.00
	1.0	.7	.04	.01	.00
		.9	.11	.09	.02
.25		.3	.21	.06	.01
	.3	.7	.17	.11	.01
		.9	.30	.14	.02
		.3	.17	.07	.00
	.5	.7	.26	.13	.04
		.9	.23	.10	.02
		.3	.17	.07	.00
	.7	.7	.21	.09	.02
		.9	.46	.26	.05
		.3	.17	.07	.01
	.9	.7	.42	.19	.05
		.9	.59	.41	.12
		.3	.22	.06	.00
	1.0	.7	.50	.20	.04
		.9	.68	.55	.21
.50		.3	.78	.56	.24
	.3	.7	.79	.65	.24
		.9	.86	.74	.29
		.3	.76	.60	.27
	.5	.7	.86	.72	.43
		.9	.92	.85	.62
		.3	.90	.68	.29
	.7	.7	.93	.80	.44
		.9	.95	.89	.63
		.3	.94	.83	.34
	.9	.7	.97	.91	.70
		.9	1.00	1.00	.91
		.3	.96	.89	.41
	1.0	.7	1.00	.96	.74
		.9	1.00	1.00	.99

Significant Three-Way Interaction from MSE ANOVA



SAMPLE SIZE BY CORRELATION BY RELIABILITY

— Figure 2 —

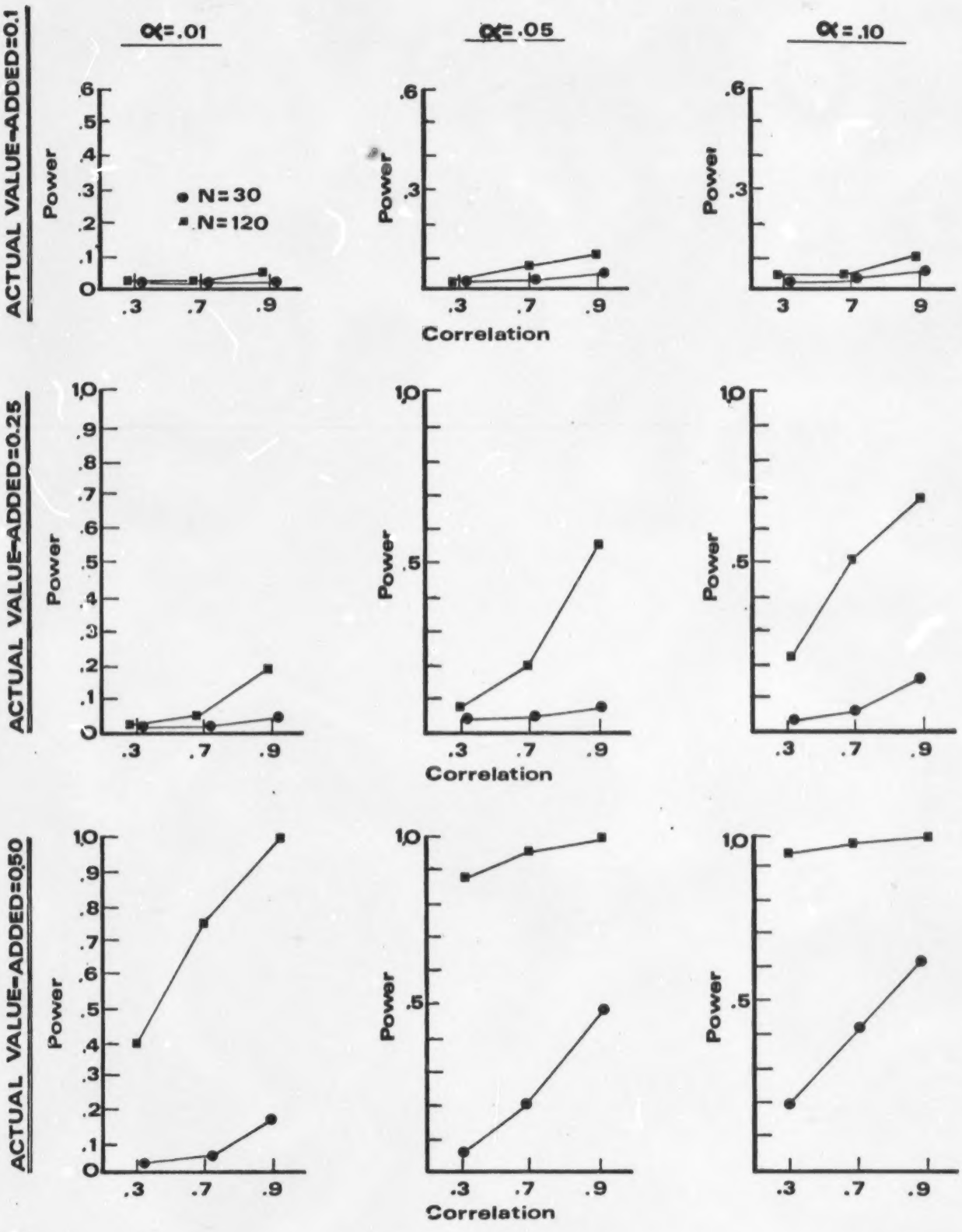


Fig. 3: Power When Test Reliability = 1.0

at least .7 or if the test reliability is .5 and the age-true score correlation is .9. Even at N = 120, power at $\alpha = .01$ is never very high unless the age-true score correlation is greater than .7, the treatment effect is large, and the test reliability is high.

Nonlinear Growth Models

One should recall that the test scores could not exceed 107, and that if a subject's score would have exceeded 107 based on the subject's age, error term or value-added, the score was set to 107. If necessary, the value-added, error term, or expected pretest-posttest maturation terms were truncated so that the test score would not be greater than the test ceiling. were also near to the intended value, 30.25. The percentage of subjects in the various treatment groups scoring at the test ceiling ranged from 4.6%-12.3% at the pretest to 34.1%-45.0% at the posttest (see Table 11). The percentage of subjects scoring at the test ceiling on the posttest increased as the size of the treatment effect increased and as the size of the standard deviation of the treatment effect increased from 4-6%-12.3% at the the pretest to 34.1%-45.0% at the posttest. This ceiling effect had the result of reducing both the mean and the variance of the value-added. On the average, the mean values-added were one-sixth to one-half of what they would have been in the absence of a test ceiling; the obtained variances were only one-half to two-thirds of what they otherwise would have been. The truncation process also reduced the test score variances, especially the posttest variances, since the ceiling effect is more pronounced at the posttest than at the pretest.

In any event, the original designations for the treatment conditions have become misleading; the mean values-added are not really a tenth, quarter, or half of the posttest standard deviation and the standard deviation of the value-added is no longer either equal to, or half of, the posttest standard deviation. In order to avoid misrepresentation, new designations will be used for the various treatment conditions. The size of the mean actual value-added (ACTUALV) will be coded 1 (smallest) to 3 (largest); the standard deviation of the value added (VALSD) will be coded with 1 (to denote the smaller value) or 2 (for the larger value).

Table 6 contains test score means and variances and Table 7 gives data concerning estimates of the value-added. Several things about these data are notable. First, as described above, the posttest vari-

Table 6

Test Score Means and Variances

Variable	Mean for Allometric Data	Mean for Negative Exponential Data
Pretest mean	80.088	85.165
Posttest mean	93.293	94.722
Pretest variance		
overall	295.323	265.847
for VALSD=1	294.660	265.776
for VALSD=2	295.986	265.919
Posttest variance		
overall	256.562	220.097
for VALSD=1	216.709	180.232
for VALSD=2	296.416	259.963

Table 7

Characteristics of Estimates of the Value - Added

Group	Mean of Actual Value-Added	Mean of Estimated Value-Added (v)		Standard Error of v		Mean bias of v		Mean Squared Error	
Allometric Data		OLS	WLS	OLS	WLS	OLS	WLS	OLS	WLS
Total Sample	2.107	1.181	0.588	4.067	3.269	-0.926	-1.518	36.204	25.083
ACTUALV=1	0.299	-0.403	-1.027	3.997	3.233	-0.702	-1.326	36.369	24.902
ACTUALV=2	1.770	0.366	-0.140	4.00	3.23	-1.404	-1.910	35.949	25.176
ACTUALV=3	4.251	3.580	2.932	4.011	3.195	-0.671	-1.319	36.293	25.170
Negative Exponential Data									
Total Sample	2.045	-0.845	-2.335	4.002	3.160	-2.890	-4.380	44.466	41.138
ACTUALV=1	0.257	-2.565	-3.947	4.065	3.233	-2.822	-4.204	45.775	40.607
ACTUALV=2	1.779	-1.029	-2.610	4.026	3.187	-2.809	-4.390	42.883	41.036
ACTUALV=3	4.098	1.060	-0.449	3.915	3.060	-3.038	-4.546	44.740	41.771

ances are smaller than the pretest variances (295.323 vs. 256.562 for the allometric data and 265.847 vs. 220.097 for the negative exponential data); this is the opposite of what one would find if there were no ceiling effects. Second, the WLS estimates of v had a smaller standard error than the OLS estimates (3.222 vs. 4.025 for the allometric data and 3.160 vs. 4.002 for the negative exponential data). Although all of the estimates of v were, on the average, negatively biased, the WLS estimates tended to exhibit more bias than the OLS estimates. All of the biases seem relatively large, considering the size of the treatment effects. Finally, MSE tended to be somewhat smaller with WLS estimates than OLS estimates.

Given the small treatment effects and relatively large biases and standard errors, is is not surprising that they power of the significance test for v is very low (see Table 8). In the case of the allometric data, power never exceeds .52 and it is very often less than .10. Power is even lower for the negative exponential data, where it never gets above .15. OLS and WLS estimates yield significance tests of comparable power.

Results of analysis of variance on mean squared error (MSE) are given in Tables 9 and 10; MSE means for the various levels of the treatment effects are given in Table 11. The main effect for sample size was highly significant for both allometric and negative exponential data and for both OLS and WLS estimates. As was the case when linear growth was assumed, an increase in sample size produced a corresponding reduction in MSE. None of the other main effects (ACTUALV, VALSD) was significant, nor were any of the interactions.

DISCUSSION

The questions of when and how to apply the value-added model are complex ones. Several conceptual and theoretical considerations need to be addressed.

First, a readily specifiable growth function is required; i.e., there must be some sort of logical developmental relationship between age and the skills being measured, at least over the duration of the program. Students should be at a similar, early stage of development when the program begins.

Table 8

Power of the Significance Test for v: Nonlinear Data

	ALLOMETRIC DATA						NEGATIVE EXPONENTIAL DATA					
	OLS			WLS			OLS			WLS		
Significance Level	.10	.05	.01	.10	.05	.01	.10	.05	.01	.10	.05	.01
N = 30	.08	.05	.01	.08	.05	.01	.04	.02	.00	.01	.00	.00
ACTUALV=1	.05	.03	.01	.04	.03	.01	.03	.02	.01	.01	.01	.00
VALSD=1	.08	.03	.01	.08	.06	.01	.04	.02	.01	.02	.01	.00
VALSD=2	.02	.02	.00	.00	.00	.00	.02	.01	.00	.00	.00	.00
ACTUALV=2	.03	.02	.00	.02	.01	.00	.04	.02	.00	.01	.00	.00
VALSD=1	.04	.02	.00	.00	.00	.00	.02	.01	.00	.01	.00	.00
VALSD=2	.02	.02	.00	.03	.01	.00	.06	.02	.00	.01	.00	.00
ACTUALV=3	.16	.10	.03	.20	.11	.02	.06	.02	.00	.02	.01	.00
VALSD=1	.21	.11	.02	.25	.12	.03	.08	.03	.00	.04	.01	.00
VALSD=2	.11	.09	.03	.14	.09	.01	.04	.01	.00	.00	.00	.00
N = 120	.15	.10	.04	.14	.10	.04	.04	.03	.01	.01	.01	.00
ACTUALV=1	.03	.02	.00	.02	.01	.00	.01	.00	.00	.00	.00	.00
VALSD=1	.04	.02	.00	.02	.01	.00	.01	.00	.00	.00	.00	.00
VALSD=2	.02	.01	.00	.01	.01	.00	.00	.00	.00	.00	.00	.00
ACTUALV=2	.10	.06	.01	.07	.04	.00	.03	.02	.01	.01	.01	.00
VALSD=1	.13	.10	.01	.10	.06	.00	.03	.02	.00	.00	.00	.00
VALSD=2	.06	.01	.01	.04	.02	.00	.03	.02	.01	.01	.01	.00
ACTUALV=3	.34	.23	.10	.35	.25	.11	.10	.08	.03	.04	.02	.01
VALSD=1	.47	.36	.17	.52	.38	.19	.15	.11	.05	.07	.04	.01
VALSD=2	.20	.10	.02	.17	.11	.03	.04	.04	.01	.00	.00	.00

Table 9

Analysis of Variance on Mean Squared Error

Allometric Data

Ordinary Least Square Estimates

Source	SS	DF	MS	F	P
SAMSIZ	574,950.188	1	574,950,188	714.55	.001
ACTUALV	40.664	2	20.332	0.03	ns
VALSD	19.822	1	19.822	0.03	ns
SAMSIZ by ACTUALV	1,776.296	2	888.148	1.104	ns
SAMSIZ by VALSD	576.306	1	576.306	0.72	ns
ACTUALV by VALSD	1,495.160	2	747.580	0.93	ns
SAMIZ by VALSD by ACTUALV	1,197.625	2	598.813	0.74	ns
Error	955,[illegible]0.938	1188	804.630	- - -	ns

Weighted Least Squares Estimates

SAMSIZ	230,688.938	1	230,688.938	566.44	.001
ACTUALV	19.033	2	9.517	0.02	ns
VALSD	640.651	1	640.651	1.57	ns
SAMSIZ by ACTUALV	1,120.300	2	560.150	1.38	ns
SAMSIZ by VALSD	6.521	1	6.521	0.02	ns
ACTUALV by VALSD	1,417.581	2	708.791	1.74	ns
SAMSIZ by VALSD by ACTUALV	1,016.498	2	508.249	1.25	ns
Error	483,822.875	1188	407.258	- - -	ns

Table 10

Analysis of Variance on Mean Squared Error

Negative Exponential Data

Ordinary Least Squares Estimates

Source of variation	SS	DF	MS	F	P
SAMSIZ	651,693.500	1	651,693.500	410.79	.001
Actualvalue-added (ACTUALV)	1,719.872	2	859.936	0.54	ns
Standard deviation of value-added (VALSD)	357.261	1	357.261	0.23	ns
SAMSIZ by ACTUALV	329.095	2	329.095	0.21	ns
SAMSIZ by VALSD	261.898	1	261.898	0.17	ns
ACTUALV by VALSD	2,162.340	2	2,162.340	1.36	ns
SAMSIZ by ACTUALV by VALSD	3,968.957	2	3,968.969	2.50	ns
Error	1,884,670.000	1188	1,586.422	- - -	- -

Weighted Least Squares Estimates

SAMSIZ	228,213.938	1	228,213.938	202.48	.001
ACTUALV	275.661	2	137.831	0.12	ns
VALSD	147.755	1	147.755	0.13	ns
SAMSIZ by ACTUALV	116.399	2	58.199	0.05	ns
SAMSIZ by VALSD	814.858	1	814.858	0.72	ns
ACTUALV by VALSD	524.165	2	262.082	0.23	ns
SAMSIZ by ACTUALV by VALSD	4,852.473	2	2,426.236	2.15	ns
Error	1,339,015.000	1188	1,127.117	- - -	- -

Table 11

Mean Values of Mean Squared
Error for Various Levels of the Treatment Effects

	ALLOMETRIC DATA		NEGATIVE EXPONENTIAL DATA	
	OLS	WLS	OLS	WLS
Total:	36.204	25.083	44.466	41.138
Size of Variance of Treatment Effect				
VALSD=1	36.330	24.350	43.919	401.785
VALSD=2	36.077	25.815	45.013	41.490
Size of Mean Treatment Effect				
ACTUALV=1	36.37	24.90	45.78	40.61
ACTUALV=2	35.95	25.17	42.88	41.03
ACTUALV=3	36.29	25.17	44.74	41.77
Sample Size				
N=30	58.09	38.95	67.77	54.93
N=120	14.31	11.22	21.16	27.35

The twin issues of content validity and construct validity are always important considerations in experimental design; when the value-added model is used, they are crucial. Haney and Gelberg (1980) and Yurchak (1979) point out that a problem in preschool program evaluation is that there is sometimes little agreement among program objectives, activities, and evaluation procedures. For example, a program might have affective objectives, activities geared toward cognitive development, and an evaluation instrument that focuses on psychomotor skills. It would obviously be difficult to draw any useful conclusions from the results of such a study. Consequently, there must be congruence among program objectives, program activities, and tests used for program evaluation.

The stable universe assumption (Kodlin & Thompson, 1958) requires that children be fairly homogeneous in age and previous educational experiences. For preschool program evaluation, Bryk and Woods (1980) suggest an age range of less than 24 months and a population with little or no formal schooling. The requirement of a stable growth rate also suggests that unusually bright younger children or slow older children should be eliminated from the analysis (Bryk and Woods, 1980). Such cases could be detected as outliers on an age by pretest score scatterplot. However, if more than a few such outliers appear, the value-added model should probably not be applied.

There should be a relatively strong linear relationship between age (or age and background variables) and test score; if there is not, a linearizing transformation should be used; The results of the second part of this study indicate that nonlinearity caused by floor or ceiling effects can result in serious problems in the estimation of the treatment effect. Among these problems are bias and inefficient estimates. Weighted least squares is not a panacea; it increased efficiency, but did so at the cost of increasing bias. A possible alternative to both OLS and WLS is tobit analysis (Tobin, 1958), a regression technique that can be applied when the dependent variable has an upper or lower limit.

The usual intuitively appealing rationale for the use of WLS is that it gives greater weight to observations at levels of the independent variable where there is relatively small variation in the dependent variable. The small variance is taken to mean that those observations are somehow more reliable than the others. In the case where there is a ceiling effect, the reduced variance could well be the result of students clustering at the ceiling. There is no reason to believe that measurement at the test ceiling is more accurate than it is elsewhere; in fact, probably the opposite is true, especially for high-achieving

students. Consequently, putting more weight on the ceiling observations served to increase the bias of the estimated treatment effect.

A related consideration is much simpler: if the test has a ceiling, the ceiling limits the possible size of an observable treatment effect. This is an especially serious problem for students scoring near the ceiling on the pretest. Even if a program did have a positive effect, it would be hard (or impossible) to detect given the test ceiling. This difficulty is not unique to the value-added model and it cannot be overcome by transformation or the use of weighted least squares. The basic stumbling block is that one can never know what the experimental results would have been if the test had a higher maximum score.

The results of this study suggest that the value-added technique does not yield powerful significance tests. A sample size of 30 is inadequate to detect many treatment effects and even a sample size of 120 is not large enough to detect very small treatment effects (0.1σ) at all, or even some larger effects ($.25\sigma$-$.50\sigma$) when the test reliability and age-pretest correlations are low. In any case, power at $\alpha = .01$ and even at $\alpha = .05$ is often unacceptably low; it is worthwhile to consider using a .10 significance level if the cost of making a type II error is high. In such cases, the additional risk of falsely rejecting the null hypothesis is probably more than offset by the increased likelihood that any existing treatment effect would actually be detected.

The results of the study also indicate that age-pretest correlation and test reliability are important determinants of power, particularly at smaller sample sizes. Consequently, care should be taken to use a reliable test that is highly correlated with age. It is entirely possible that there is some sample size beyond which age-pretest correlation and test reliability are essentially immaterial in affecting power. That question and also the interactions between sample size and test reliability and sample size and age-test score correlations could be fruitful areas for future research.

It is apparent that sample size is a very important consideration in the design of studies employing the value-added model. The data from this study suggest that for most purposes, a minimum acceptable sample size (from the standpoint of power) would not be less than 120. This limits the situations in which the model could be successfully applied. Large school systems should have enough preschool students to conduct a reasonably powerful analysis. State educational agencies might also be

able to achieve large enough sample sizes by aggregating comparable data from several school districts.

Given an adequately specified growth function, a sufficiently large sample size, and a reliable, valid test which is both well-correlated with age and free from ceiling and floor effects, the value-added model provides a promising alternative to the use of other quasi-experimental designs. The model may be applicable to program evaluation data when the variable being measured is a function of age or time. Hence, the strategy is uniquely appropriate for evaluating early childhood programs, such as those funded under Title I of the Elementary and Secondary Education Act.

Existing evaluation models are often difficult to apply to preschool programs and the existing evaluation models are often difficult to apply to preschool programs and the value-added model could be an important tool for the evaluator of such programs. The value-added model has the dual advantages of not requiring a control group and of allowing for great flexibility in the pretest-posttest time interval; if several groups are being compares, it is not necessary that the length of time between the pretest and posttest be identical for all of the groups. The results of program evaluations can serve as valuable guides in setting future educational policies. It is highly desirable to know what types of programs are most beneficial, especially when financial resources are very limited.

The value-added model may be applied in other areas in which there is a fairly definite starting point and a known pattern of growth over time. One such area would be language acquisition. Another field in which the value-added model could prove helpful is medicine. If there is a relationship between time and the development of a pathological condition, the model could be used to study the effects of either an ameliorative treatment, or a factor such as smoking that could aggravate the condition. Random assignment of subjects to experimental groups is often not possible in this type of research and the value-added model could prove to be a useful alternative design.

Application of the value-added model to non-preschool settings would be a useful area for future research. More work also needs to be done in determining sample sizes that will yield adequately powerful significance tests in cases where background variables such as race, sex, or socioeconomic status have been used along with age or time in order to yield more precise estimates of growth rate for particular groups of subjects.

This study has several limitations, most of which relate to the degree to which the results can be generalized to other situations. Such problems represent a basic shortcoming of the Monte Carlo technique. In the linear growth model section, the variance of the value-added was set equal to the posttest variance, which was in turn set equal to the pretest variance. Since actual value-added model data were not available, it is not known whether these assumptions are realistic. It is also possible that different results would have been obtained had the mean or variance of the value-added assumed other values. Such questions could be explored in future research.

In the case of nonlinear growth models, there is the question of whether the means and variances of the value-added correspond to what researchers might encounter in practice. The estimates of power would not easily generalize to other nonlinear growth models, especially if the percentages of subjects scoring at the ceiling decisions.

In conclusion, this study has demonstrated some of the advantages and disadvantages of value-added analysis. Given a fairly large sample size and an accurate measuring device, the technique provides a means of studying phenomena which were previously difficult or impossible to adequately explore.

REFERENCES

Bryk, A. S., & Weisberg, H. I. A new approach to analyzing quasi-experimental data: Value-added analysis. Proceedings of the Social Statistics Section, American Statistical Association, 1974.

Bryk, A. S., & Weisberg, H. I. Value-added analysis: A dynamic approach to the estimation of treatment effects. Journal of Educational Statistics, 1976, 1, 127-155.

Bryk, A. S., & Weisberg, H. I. Use of the nonequivalent control group design when subjects are growing. Psychological Bulletin, 1977, 84, 950-962.

Bryk, A. S., Strenio, J. F., & Weisberg, H. I. A method of estimating treatment effects when individuals are growing. Journal of Educational Statistics, 1980, 5, 5-34.

Bryk, A., and Woods, E. An introduction to the value-added model and its use in short term impact assessment. Huron Institute: Cambridge MA, 1980.

Cook, T. D., & Campbell, D. T. Quasi-experimentation: Design & analysis issues for field settings. Chicago: Rand McNally, 1979.

Haney, W. Short-term impact evaluations of early childhood Title I programs (Draft). Cambridge, Mass.: Huron Institute, 1980.

Haney, W., & Gelberg, W. Assessment in early childhood education (Draft). Cambridge, Mass.: Huron Institute, 1980.

Haney, W., & Gelberg, W. Summative evaluation of early childhood Title I programs: Issues of feasibility and instrumentation. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Huron Institute. Report to participating LEAs re: Value-added strategy. Unpublished, 1/24/80.

IMSL, Inc. International Mathematical and Statistical Library reference manual. Houston: IMSL, 1980.

Kodlin, P., & Thompson, D. J. An appraisal of the longitudinal approach to studies of growth and development. Monographs of the Society for Research in Child Development, 23(1), 1958. Miller, R. G. The jackknife - a review. Biometrika, 1974, 61(1), 1-15.

Mosteller, F., & Tukey, J. W. Data analysis and regression. Reading, Mass.: Addison-Wesley, 1977.

Strenio, J. F. An individual growth model perspective for evaluation evaluating educational programs. Qualifying Paper, Harvard University Department of Statistics, 1977.

Tobin, J. Estimation of relationships for limited dependent variables. Econometrica, 1958, 26, 24-36.

Yurchak, M. J. Current implementation and evaluation practices in early childhood Title I programs. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.